EMERGING TEXT MINING TECHNIQUESIN BIG DATA ANALYTICS

D.ARPITHA RANI, Assistant Professor Loyola Academy Degree and P.G College, Secunderabad
S. SARASHRI, Assistant Professor Loyola Academy Degree and P.G College, Secunderabad
arpitharani8@gmail.com; sarashri1817@gmail.com

ABSTRACT:
This paper summarizes research on Big data analytics which was growing rapidly with the flood of data traffic exchanged day by day. As the increase of vast data demands better data-driven decisions. In view of commercial effect and to overcome competition we are looking forward to the best decision making to manage this we need to go through the optimization of huge sets of data i.e., datasets. Here the idea in this paper is to emerge Text mining techniques with big data which is a powerful tool to analyze textual data and extract knowledge-based information for decision-making. Text mining also identifies various patterns and correlations present in the data. Mining data of various types lead to different challenges these days there are many tools that are developed to accept various forms and size of data related to the digital world. Text mining techniques such as word association analysis, text clustering, word-level analysis, topic modeling, and information retrieval, sentiment analysis, advanced techniques like how fraud detection can be identified is specified in this paper.

KEYWORDS: Big data, Text mining, text clustering, decision making, fraud detection.

INTRODUCTION

In recent times BIGDATA plays a specific role in accepting data based on four v's that is velocity, variety, volume and veracity from diverse source of platforms in different design of data like structured, semi-structured and unstructured .big data aids businesses in utilizing their data to find new opportunities. This leads to more profitable business decisions, more efficient operations, and happier customers. Cost reduction is one of the many benefits that big data and advanced analytics provide for businesses with the ultimate goal of real-time analysis, big data analytics will increasingly put a priority on data freshness, enabling more informed decisions and greater competitiveness. Text databases are rapidly growing as a result of the large amount of information that is available electronically. Today, more than 80% of knowledge is unstructured or only loosely organized. The growing volume of text data renders outdated information retrieval methods ineffective.

Text mining focuses on information discovery from huge corpora of unstructured text, which is essential for dealing with a variety of natural language processing tasks including such text representation models, indexing and classification, topic analysis, semantic similarity search, explaining patterns and subjects of interest (also known as descriptive text mining), sentiment analysis and opinion mining, text summarization, chatbot or digital assistant generation, and so on. Data recovery, data mining, AI, statistics, machine learning, and computational linguistics are all components of the multidisciplinary discipline of text mining.

Text mining techniques

Word association analysis

Following a brief introduction, this section discusses extracting word associations from text data. The paradigmatic relation and the syntagmatic relation are the two most fundamental word relationships. These two types of relationships are fundamental and can be generalised to include fundamental connections between elements in any order. Additionally, they can be used to generalise the relationships between any elements in a language, including words and even complicated sentences. If elements A and B may be substituted for one another, they are said to have a paradigmatic relationship.

Thus, the two terms belong to the same semantic class or syntactic category. In most cases, they can be swapped out for one another without impairing the reader's comprehension, so the statement would still be understood. On the other hand, the two terms can be combined with one another in a syntagmatical relation. As a result of their semantic similarity, the elements A and B can be combined with one another in a sentence. However, generally speaking, they cannot be substituted for one another because the statement would lose all meaning. Another way to look at the relationships is to think of them as: (1) relationships that occur in comparable places in regard to the neighbours in the sequence (paradigmatic relationship); or (2) relationships involving co-occurring elements that frequently appear in the same sequence (syntagmatical relation). These two fundamental word relationships are complementary. One can presume that words with high context similarity also have paradigmatic relation while trying to uncover it. Therefore, each word's context similarity must be calculated. One must look for words with high co-occurrences but relatively low individual occurrences in order to find syntagmatic link. The reason for this is that certain terms frequently appear together. One must count the number of times two words appear together in a context in order to calculate the syntagmatic relationship (a sentence, a paragraph, or even a document). The co-occurrences and their individual occurrences should then be compared. Since words that are syntagmatically related to one another typically have a close relationship, both paradigmatic and syntagmatic relations are closely related. They appear to be related to the same word, which implies that the two relationships can be discovered together. Some of the statistical techniques used to find those kinds of relationships are introduced in the following sections.

Text clustering

The goal of text clustering is to arrange a collection of unlabeled texts into groups where texts within each group are more similar to one another than to texts in other groups. Text clustering algorithms analyze the text to see if there are any natural clusters (groups) in the data. a procedure that uses different clustering algorithms to autonomously classify text content into groups. When clustering is done in a top-down and bottom-up manner, comparable terms or patterns are arranged and extracted from various papers. As a consequence, separate divisions, referred to as clusters, are produced, and each cluster contains a number of documents. The content of every document in a single cluster is remarkably similar, whereas the content of documents in various clusters is distinct, improving the quality of clustering. A basic algorithm for clustering texts tracks each document's subjects and assigns weights based on how well the papers fall into each cluster.

Sentiment analysis : The goal of sentiment analysis (SA) is to glean sentiments, feelings, or views from texts that have been made public by various data sources, such as SNs. In-depth research on SA's primary tools and approaches is presented in this article. analysis of feelings. In the framework of sentiment analysis, we look into image to text mapping.

The majority of sentiment analysis research conducted in the past uses text data. Recent studies (Yu et al., 2016; You et al., 2015; Wang et al., 2016) have put a lot of attention on sentiment analysis in pictures and videos.

Visual sentiment analysis research is conducted in two ways: first, using hand-crafted features, and second, using features produced automatically. A huge number of photos can be used to automatically build robust characteristics using deep learning algorithms (Jindal and Singh, 2015). Word representations and capsule networks are a fascinating area for sentiment analysis in NLP applications.

K-Means AlgorithmsThe k-imply technique divides the facts set into kclusters, wherein each cluster issubjected to be represented through the imply of points; referred to as the centroid. A two-steprepetitive technique is hired for theutility of the set of rules:

(1) Assigningevery factor to the closest centroid. (2) Evaluating the centroids for a recentlydeveloped group. The technique is ended while the cluster centroid involves aconstant price. The k-imply set of rules has an in depth utility because of itsdirect parallelization. Furthermore, the order of respective facts does now no longer have an effect on the k-imply set of rules which attributes the numerical traits to it. It is needed tomention the most price of kat the start of the technique. The representationof the cluster is made through the k-medoid set of rules that chooses the item adjoiningthe middle of the cluster. Though, the choice of the kobjects is carried out randomly inthe set of rules. The decided on gadgets assist to decide the distance. A cluster isformed on the idea of the closest item to k, while the opposite gadgets collect theposition of krecursively until the specified excellent of the cluster is achieved

Topic modeling: An abstract "theme" that appears in a group of documents can be found using a topic model, a form of statistical model. To find latent semantic structures in a text body, topic modelling is a popular text-mining tool.

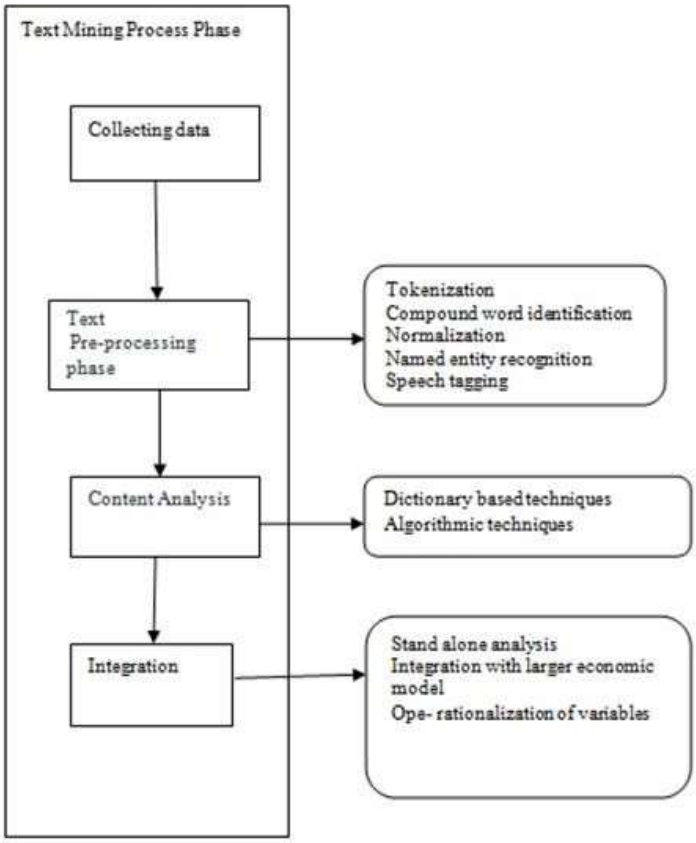

Figure:1

INFORMATION RETRIEVAL (IR)

A software programme used for information retrieval (IR), particularly for text-based information, can be defined as one that organizes, stores, retrieves, and evaluates information from document repositories. Although the system helps users locate the data they need, it does not clearly give users the answers to their inquiries. It discloses the existence and whereabouts of any documents that may contain the necessary information. Relevant documents are those that satisfy the user's request. A great IR system will only return pertinent materials.

Summarizing study

Text summary condenses information in an effort to make it simpler for readers to recognize and comprehend related source materials. Significant work has been done in recent years to develop and test various techniques. for the various domains. Text summarization, or more specifically automated text

summarization, is the process by which a computer creates a condensed version of the original text (or group of texts), while maintaining the majority of the information available in the original text. Although some information may be lost during this process, it is still very helpful for data compression. For instance, there are many research resources available in the biomedical field, and summarizing this information is quite helpful to researchers.

Trial and Error Analysis:

Assembling unstructured information from sources available in a range of document formats, such as plain text, web pages, PDF records, etc. Data consistency is identified and removed from the data by pre-processing and data cleansing operations. The data cleansing procedure ensures that only authentic text is recorded and is used to get rid of stop words stemming (the process of figuring out a word's root and indexing the data).The data set is reviewed and further cleaned using processing and controlling operations. In the Management Information System, pattern analysis is used. For a powerful and practical decision-making process and trend analysis, information processed in the aforementioned processes is used to extract pertinent and significant data.

Text Analysis methods on Social Media

Application of Personality Representation Based on Facebook Data Extracted Features. In light of the insightful research findings, the categorization methods and their applications were thoroughly examined. The research study used 250 user instances from Facebook, drawn from a sample of 10,000 status changes provided by the My Personality project [53].The following two related goals of the study are interconnected: Knowing the relevant personality-correlated indications that Facebook uses to obliquely or overtly convey user data, as well as determining the viability of demonstrating prognostic character in order to enable emerging intelligent systems, are two prerequisites. In order to improve the output of the classifiers being evaluated, the study focused on promoting relevant characteristics in models.

Data is being used across a wide range of fields, some for applications and some for scholarly research [55]. This section presents updated developments pertaining to twitter data. The "Text Mining" process is started by the collecting of documents from diverse sources. By using a text mining tool to locate a specific document, its character sets and formatting are checked as part of its pre-processing [56]. The document would then be monitored by a text analysis step. It is referred to as "Text analysis" when semantic analysis is applied to extract reliable information from text. Text analysis methods are widely available on the market. As long as it serves the organization's objectives, professionals can combine different methodologies. Unless they have enough data, researchers frequently use the same text analysis methods.

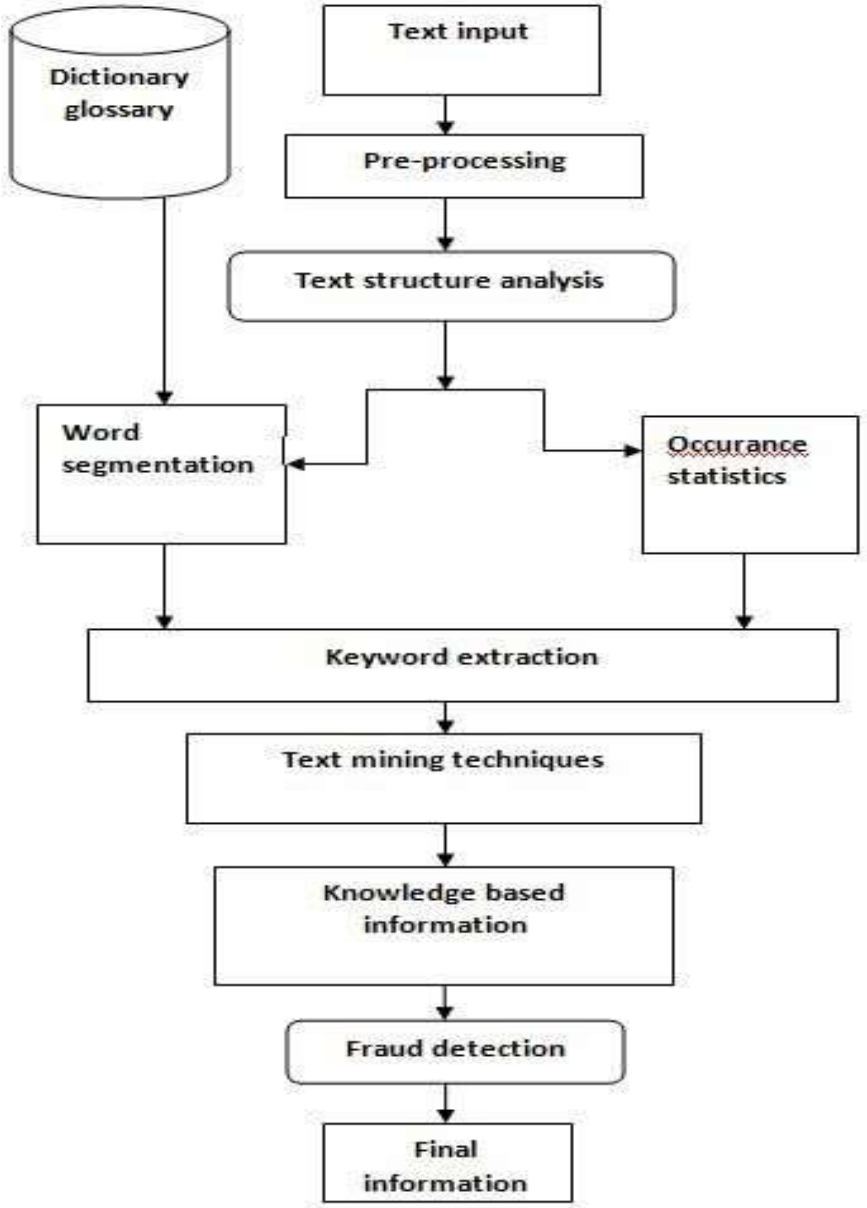

Figure 2

Fraud Detection

Text mining has helped uncover fraud cases and help entire insurance companies, financial companies, and social media. Businesses can now use text mining to process claims much faster without falling for fraudulent claims. One very common use of text mining in relation to fraud detection is spam email classification (also known as spam filtering). The classification problem is classifying spam and non-spam emails. Another application is the classification of fake reviews for all kinds of products. Classification of text into specific domains. For example, comparing spam emails to non-spam emails, or detecting sexually explicit content. Text clustering for automaticallya hard and fast of documents.

***Sentiment analysis*:** to identify and extract subjective information in documents. To find out what your customers are saying about your business when they use social media. Well, the only limit is your imagination.

Document summaries that automatically provide the most important points of the original document. This is especially suitable for summarizing messages. Learning relationships between named entities. Here is an interesting article: Identifying Semantic Relationships between Named Entities from Chinese Text .Another subtask of 7-NLP is POS or Part of Speech. This activity matches parts of speech such as nouns, adjectives, and verbs to words in your text based on their context and relationship to adjacent words. Another important task of NLP is resolving cross- references. Understanding and disambiguating references to multiple entities in text.

A text mining system takes as input to collect financial statements. The first step in detecting fraudulent financial reporting is to collect financial statements from both types of organizations (fraudulent or no fraudulent).
A company with a fraudulent history can be identified by analyzing his AAER issued by the SEC.Dataset must include non-fraudulent annual financial statements for each fraudulent organization. A non-fraud organization must be the same size (based on assets or sales) as the fraud organization. The second step is preprocessing, which extracts qualitative descriptions from financial statements and organizes them into documents. This is because documents are the basic unit of analysis in text mining. During preprocessing, convert words present in all documents to lower case so that the corpus (a collection of documents) does not contain the same words such as "Legal" and "Legal" as different words is needed. I need to remove all punctuation from the corpus, followed by all digits. The input to the classifier should be text only. Stop words such as articles (a, the, etc.), conjunctions (but, and, etc.), and prepositions (on, in, etc.) should also be removed during preprocessing document. The account space does not require stemming. This is because inflected terms can have different meanings. Because text mining considers a sentence as a collection of words and can change the order of words without affecting the analysis results, it can ignore the syntactic structure of sentences in order to process the text in an efficient way. However, we need to retain information about the occurrences of each word. This unordered collection of his words is known as the "word bag". The word bag approach uses each word occurrence as his feature to train the classifier. A "word bag" model represents each document with a vector of the number of words that appear in the document. The vectors associated with each document are compared to typical vectors associated with a particular class (fraud or non-fraud). Documents with similar vectors are considered similar in content, but otherwise dissimilar .The vector space generated above is used in the next step to classify organizations as fraudulent or non-fraudulent. We recommend using support vector machines because SVM creates a hyper plane in the feature space that best classifies fraudulent and non-fraudulent financial reports. The SVM takes a set of input data and, given any input, predicts which of two possible classes (fraud or non-fraud) will form her output. Given a set of training examples, each identified as belonging to one of two categories, the SVM training algorithm builds a model that assigns new examples to either category. SVM is a supervised machine learning technique, so it learns from the feature space of both cheating and non-cheating examples present in the training set. Once this method is learned, it will be able to correctly classify fraudulent and non-fraudulent organizations in the test data set. The accuracy of classification should be evaluated using metrics such as accuracy, precision, recall (sensitivity of binary classification), F-score, and purity.

References

[1] Talabis, M.R.M.; McPherson, R.; Miyamoto, I.; Martin, J.L.; Kaye, D. Security and text mining. In InformationSecurity Analytics; Talabis, M.R.M., McPherson, R., Miyamoto, I., Martin, J.L., Kaye, D., Eds.; Elsevier: Amsterdam, The Netherlands, 2015; pp. 123–150, doi:10.1016/B978-0-12-800207-0.00006-x.
[2] Hearst, M.A. Text Data Mining. In The Oxford Handbook of Computational Linguistics; Mitkov, R.,

Ed.; OxfordUniversity Press: Oxford, UK, 2005; pp. 616–662, doi:10.1093/oxfordhb/9780199276349.013.0034.
[3] Dumais, S. Using SVMs for text categorization, Microsoft research. IEEE Intell. Syst. Mag. 1998,13, 18–28. [4]Guduru, N. Text Mining with Support Vector Machines and Non-Negative Matrix Factorization Algorithms. Ph.D. Thesis, University of Rhodes Island, Rhodes Island, Greece, 2006.
[5]Bholat, D.; Hansen, S.; Santos, P.; Schonhardt-Bailey, C. CCBS Handbook No. 33, Text Mining For CentralBanks;Bank of England: London, UK, 2015.
[6]L. Hirschman, R. Gaizauskas (2001), "Natural language question answering: the view from here", NaturalLanguage Engineering 7. Cambridge University Press. (online reading: http://www.loria.fr/~gardent/ applicationsTAL/papers/jnle-qa.pdf)
[7]OpenEphyra (2011): https://mu.lti.cs.cmu.edu/trac/Ephyra/wiki/ OpenEphyra (accessed 5 January 2016)]
[8]N. Schlaefer, P. Gieselmann, and G. Sautter (2006). "The Ephyra QA system". 2006 Text Retrieval Conference(TREC).
[9] YodaQA (2015): http://ailao.eu/yodaqa/ (accessed 5 January 2016)
[10] P. Baudis (2015) "YodaQA: A Modular Question Answering System Pipeline". POSTER 2015 — 19thInternational Student Conference on Electrical Engineering. (online reading: http://ailao.eu/yodaqa/yodaqaposter2015.pdf)
[11] DL4J (2015): http://deeplearning4j.org/textanalysis.html (accessed 16 December 2015)
[12] Google – Word2vec (2013): http://arxiv.org/pdf/1301.3781.pdf (accessed 20 December 2015)
[13] D. Lazer, R. Kennedy, G. King, and A. Vespignani (2014). "Big data. The parable of Google Flu: traps in bigdata analysis." Science, 343(6176).
[14]. D. Boyd, and K. Crawford (2011). "Six Provocations for Big Data". A Decade in Internet Time: Symposium onthe Dynamics of the Internet and Society. (Available at SSRN: http://ssrn.com/abstract=1926431 or http:// dx.doi.org/10.2139/ssrn.1926431)
[15]. A. Moreno, and E. Moro (2015). "Big data versus small data: the case of 'gripe' (flu) in Spanish". Procedia,Social and Behavioral Sciences, 198.
[16].B. Liu (2012). Sentiment Analysis and Opinion Mining. Morgan and Claypool. Chicago.
[17]. D. Garcia, A. Garas, and F. Schweitzer (2012). "Positive words carry less information than negative words".EPJ Data Science, 1:3. (online reading: http://www.epjdatascience.com/content/1/1/3)
[18]. M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede (2011).
[19] Hoogs Bethany, Thomas Kiehl, Christina Lacomb and DenizSenturk (2007). A Genetic Algorithm Approach to Detecting Temporal Patterns Indicative Of Financial Statement Fraud, Intelligent systems in accounting finance and management 2007; 15: 41 – 56, John Wiley & Sons, USA, available at: www.interscience.wiley.com.
[20] BelinnaBai, Jerome yen, Xiaoguang Yang, False Financial Statements: Characteristics of china listed companiesand CART Detection Approach, International Journal of Information Technology and DecisionMaking , Vol. 7, No. 2(2008), 339 – 359.
[21] Ibrahim H. , Ali H. "The use of data mining techniques in detecting fraudulent financial statements: An application on manufacturing firms", The journal of faculty of economics and administrative sciences, (2009) Vol.14, No. 2 pp. 157 – 170.

